Assistant Commissioner for Patents
Washington, D.C. 20231
Sir:

ATTORNEY DOCKET: Y0999-046 (8728-252)
Date: April 2, 1999
Express Mail Label: EL192803193US
Date of Deposit: April 2, 1999

A

Transmitted herewith for filing is the Patent Application of:

Inventors: Raimo Bakis, Ellen M. Eide

For: SYSTEM AND METHOD FOR RESCORING N-BEST HYPOTHESES OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM

Enclosed are: [X] 18 sheets of specification; [X] 1 sheet(s)of Abstract; [X] 10 sheet(s) of claims; [X] 3 sheet(s) of drawing(s);

[X] An assignment of the invention to International Business Machines Corporation with Recordation Form.

[X] Declaration and Power of Attorney.

[ ] A certified copy of a ______________ application, from which priority under Title 35 USC §119 is claimed.

[X] Associate Power of Attorney.

The filing fee has been calculated as shown below:

	(Col. 1)	(Col. 2)
FOR:	NO. FILED	NO. EXTRA
BASIC FEE	▨	▨
TOTAL CLAIMS	22 -20 =	2
INDEP CLAIMS	3 -3 =	0
MULTIPLE DEPENDENT CLAIMS PRESENTED		

If the difference in Col. 1 is less than zero, enter "0" in Col. 2.

OTHER THAN A SMALL ENTITY

RATE	FEE
▨	$760.00
X $18 =	$36.00
X $78 =	0
+ 260 =	
TOTAL	$ 796.00

[ ] A check in the amount of $_____ to cover the [ ] filing fee(s), [ ] recording fee is enclosed.

[X] Please charge my Deposit Account No. 50-0510/IBM (Yorktown Heights) in the amount of $796.00.

[X] The Commissioner is hereby authorized to charge payment of the following fees associated with this communication or credit any overpayment to Deposit Account No. 50-0510/IBM (Yorktown Heights). A duplicate copy of this sheet is enclosed.

[X] Any additional filing fees required under 37 CFR 1.16.

[X] Any patent application processing fees under 35 CFR 1.17.

Respectfully submitted,

By: ______________
Frank V. DeRosa
Registration No. 43,584

Please address all correspondence to:
F. CHAU & ASSOCIATES, LLP
1900 Hempstead Tpke., Suite 501
East Meadow, NY 11554
Tel: (516) 357-0091
Fax: (516) 357-0092

Attorney for:
IBM Corporation
Intellectual Property Law Dept.
P.O. Box 218
Yorktown Heights, NY 10598

CERTIFICATION UNDER 37 C.F.R. §1.10

I hereby certify that this Application transmittal and documents referred to as enclosed are being deposited with the United States Postal Service on this date April 2, 1999 in an envelope as "Express Mail Post Office to Addressee" Mailing Label Number EL192803193US addressed to: Assistant Commissioner for Patents, Box Patent Application, Washington, D.C. 20231.

Frank V. DeRosa

Please type a plus sign (+) inside this box → ☒ +

PTO/SB/05 (4/98)
Approved for use through 09/30/2000. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 C.F.R. § 1.53(b))

Attorney Docket No.	Y0999-046 (8728-252)
First Inventor or Application Identifier	Bakis
Title	SYSTEM AND METHOD FOR RESCORING N-BEST
Express Mail Label No.	EL192803193US

APPLICATION ELEMENTS
See MPEP chapter 600 concerning utility patent application contents.

ADDRESS TO: Assistant Commissioner for Patents
Box Patent Application
Washington, DC 20231

1. ☒ * Fee Transmittal Form *(e.g., PTO/SB/17)*
(Submit an original and a duplicate for fee processing)
2. ☒ Specification [*Total Pages* 29]
(preferred arrangement set forth below)
- Descriptive title of the Invention
- Cross References to Related Applications
- Statement Regarding Fed sponsored R & D
- Reference to Microfiche Appendix
- Background of the Invention
- Brief Summary of the Invention
- Brief Description of the Drawings (*if filed*)
- Detailed Description
- Claim(s)
- Abstract of the Disclosure

3. ☒ Drawing(s) *(35 U.S.C. 113)* [*Total Sheets* 3]
4. Oath or Declaration [*Total Pages* 2]
   a. ☒ Newly executed (original or copy)
   b. □ Copy from a prior application (37 C.F.R. § 1.63(d))
   (for continuation/divisional with Box 16 completed)
      i. □ DELETION OF INVENTOR(S)
      Signed statement attached deleting inventor(s) named in the prior application, see 37 C.F.R. §§ 1.63(d)(2) and 1.33(b).

**NOTE FOR ITEMS 1 & 13: IN ORDER TO BE ENTITLED TO PAY SMALL ENTITY FEES, A SMALL ENTITY STATEMENT IS REQUIRED (37 C.F.R. § 1.27), EXCEPT IF ONE FILED IN A PRIOR APPLICATION IS RELIED UPON (37 C.F.R. § 1.28).*

5. □ Microfiche Computer Program *(Appendix)*
6. Nucleotide and/or Amino Acid Sequence Submission
(*if applicable, all necessary*)
   a. □ Computer Readable Copy
   b. □ Paper Copy (identical to computer copy)
   c. □ Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

7. ☒ Assignment Papers (cover sheet & document(s))
8. □ 37 C.F.R.§3.73(b) Statement *(when there is an assignee)* ☒ Power of Attorney
9. □ English Translation Document *(if applicable)*
10. ☒ Information Disclosure Statement (IDS)/PTO-1449 ☒ Copies of IDS Citations
11. □ Preliminary Amendment
12. ☒ Return Receipt Postcard (MPEP 503)
(Should be specifically itemized)
13. □ * Small Entity Statement(s) *(PTO/SB/09-12)* □ Statement filed in prior application, Status still proper and desired
14. □ Certified Copy of Priority Document(s)
(if foreign priority is claimed)
15. ☒ Other: Associate Power of Attorney

16. If a CONTINUING APPLICATION, *check appropriate box, and supply the requisite information below and in a preliminary amendment:*
□ Continuation □ Divisional □ Continuation-in-part (CIP) of prior application No: ______/______
Prior application information: Examiner ____________ *Group / Art Unit:* ____________
For CONTINUATION or DIVISIONAL APPS only: The entire disclosure of the prior application, from which an oath or declaration is supplied under Box 4b, is considered a part of the disclosure of the accompanying continuation or divisional application and is hereby incorporated by reference. The incorporation can only be relied upon when a portion has been inadvertently omitted from the submitted application parts.

17. CORRESPONDENCE ADDRESS

□ *Customer Number or Bar Code Label* *(Insert Customer No. or Attach bar code label here)* or ☒ *Correspondence address below*

Name	Frank V. DeRosa				
Address	F. Chau & Associates, LLP 1900 Hempstead Turnpike, Suite 501				
City	East Meadow	*State*	New York	*Zip Code*	11554
Country	USA	*Telephone*	516-357-0091	*Fax*	516-357-0092

Name (Print/Type)	Frank V. DeRosa	*Registration No. (Attorney/Agent)*	43,584
Signature	Frank DeRosa	*Date*	4/2/99

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

SYSTEM AND METHOD FOR RESCORING N-BEST HYPOTHESES OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM

GOVERNMENT LICENSE RIGHTS

This invention was developed under United States Government ARPA Contract No. MDA 972-97-C0012. The United States Government has certain rights to the invention.

BACKGROUND

1. Technical Field:

The present invention relates generally to speech recognition and, more particularly, to a system and method for rescoring N-best hypotheses output from an automatic speech recognition system by utilizing an independently derived text-to-speech (TTS) system to generate a synthetic waveform for each N-best hypothesis and comparing each synthetic waveform with the original speech waveform to select the final system output.

2. Description of Related Art:

A common technique which is utilized in speech recognition is to first produce a list of the N most-likely ("N-best") hypotheses for each utterance and then rescore each of the N-best hypotheses using one or more knowledge sources not necessarily modeled by the speech recognition system which produced the N-best hypotheses. Advantageously, this "N-best rescoring" method enables additional knowledge sources to be brought to bear on the recognition task without having to integrate such sources into the initial decoding system.

One such "N-best rescoring" method is disclosed in "An Articulatory-Like Speech Production Model with Controlled Use of Prior Knowledge" by R. Bakis, Frontiers in Speech, CD-Rom, 1993. With this method, an articulatory model which generates acoustic vectors (not speech waveforms) given a phonetic transcription is utilized to produce acoustics against which the original speech may be compared. Other "rescoring" methods are known to those skilled in the art.

As is understood by those skilled in the art, the techniques utilized for speech recognition and speech synthesis are inherently related. Consequently, increased knowledge and understanding and subsequent improvements for one technique can have profound implications for the other. Due to the recent advances in text-to-speech (TTS) systems which have enabled high quality synthesis, it is to be appreciated that a TTS system can sufficiently provide a source of knowledge about what the speech signal associated with each of the N-hypothesis would look like. Currently, there exists no known systems or methods which utilize a TTS system for rescoring N-best hypotheses. Therefore, based on the similarities between speech recognition and speech synthesis, it is desirable to employ a TTS system as a knowledge source for use in rescoring N-best hypotheses.

SUMMARY OF THE INVENTION

The present invention is directed to a system and method for rescoring N-best hypotheses of an automatic speech recognition system, wherein the N-best hypotheses comprise the N most likely text sequences of a decoded original waveform. In one aspect of the present invention, a method for rescoring N-best hypotheses comprises the steps of:

generating a synthetic waveform for each of the N text sequences;

comparing each synthetic waveform with the original waveform to determine the synthetic waveform that is closest to the original waveform; and

selecting for output the text sequence corresponding to the synthetic waveform determined to be closest to the original waveform.

In another aspect of the present invention, in order to compare the original and synthetic waveforms, each is transformed into a set of feature vectors using the same feature analysis process.

In another aspect of the present invention, the original and each of the synthetic waveforms representing the Nth hypotheses are compared on a phoneme-by-phoneme basis by segmenting (aligning) the stream of feature vectors into contiguous regions, each region representing the physical representation of one phoneme in the phonetic expansion of the hypothesized text sequence.

In another aspect of the present invention, an automatic speech recognition system comprises:

a decoder for decoding an original waveform of acoustic utterances to produce N text sequences, the N text sequences representing N-best hypotheses of the decoded original waveform;

a waveform generator for generating a synthetic waveform for each of the N text sequences; and

a comparator for comparing each synthetic waveform with the original waveform to rescore the N-best hypotheses.

Advantageously, by comparing the synthetic waveforms (for each of the N most-likely text sequences) to the original waveform, one can incorporate the body of knowledge and understanding required to build the synthesis model into the N-best framework for rescoring the top N hypotheses.

These and other aspects, features and advantages of the present invention will be described and become apparent from the following detailed description of preferred embodiments, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block/flow diagram of a system/method for rescoring N-best hypotheses in accordance with an embodiment of the present invention; and

Figs. 2A and 2B comprise a detailed flow diagram of a method for rescoring N-best hypotheses in accordance with one aspect of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

It is to be understood that the system and method described herein may be implemented in various forms of hardware, software, firmware, special purpose microprocessors, or a combination thereof. Preferably, the present invention is implemented in software as an application program tangibly embodied on a program storage device. The application program may be uploaded to, and executed by, a machine having any suitable and preferred microprocessor architecture. Preferably, the machine is implemented on a computer platform having hardware such as one or more central processing units (CPU), a random access memory (RAM), and input/output (I/O) interface(s). The computer platform also includes an operating system and microinstruction code. The various processes and functions described herein may either be part of the microinstruction code or part of the application program (or a combination thereof) which is executed via the operating system. In addition, various other peripheral devices may be connected to the computer platform such as an additional data storage device and a printing device.

It is to be further understood that, because some of the constituent system components and method steps depicted in the accompanying Figures are preferably implemented as software modules, the actual connections between the system components (or the process steps) may differ depending upon the manner in which the present invention is programmed. Given the teachings herein, one of ordinary skill in the related art will be able to contemplate these and similar implementations or configurations of the present system and method.

Referring now to Fig. 1, a block diagram illustrates a system for rescoring N-best hypotheses of an automatic speech recognition system in accordance with an embodiment of the present invention. It is to be understood that the diagram depicted in Fig. 1 can also be considered a general flow diagram of a method for rescoring N-best hypotheses in accordance with the present invention. The system 100 includes a feature analysis module 101 which receives and digitizes input speech waveforms (spoken utterances), and transforms the digitized input waveforms into a set of feature vectors on a frame-by-frame basis using feature extraction techniques known by those skilled in the art. Typically, the feature extraction process involves computing spectral or cepstral components and corresponding dynamics such as first and second derivatives. Preferably, the feature analysis module 101 operates by first producing a 24-dimensional cepstra feature vector for every 10ms of the input waveform, splicing nine frames together (i.e., concatenating the four frames to the left and four frames to the right of the current frame) to augment the current vector of cepstra, and then reducing each augmented cepstral vector to a 60-dimensional feature vector using linear discriminant analysis. The input (original) waveform feature vectors are then stored for subsequent processing as discussed below.

The original waveform feature vectors are then decoded by a speech recognition system 102 having trained acoustic prototypes to recognize and transcribe the spoken words of the original waveform. In particular, the speech recognition system 102 is configured to generate N-best hypotheses 103 (i.e., the N most-likely text sequences (transcriptions) of the spoken utterances). It is to be understood that any conventional technique may be employed in the speech recognition system 102 for generating the N-best hypotheses such as the method disclosed in "The N-Best Algorithm: An Efficient and Exact Procedure For Finding the N Most Likely Sentence Hypotheses" by Schwartz, et al., pp. 81-84. Proc. ICASSP, 1990.

The N-best hypotheses 103 are input to a text-to-speech system (TTS) 104 to generate a set of N synthetic waveforms 105, each synthetic waveform being a text sequence corresponding to one of the N-best hypotheses 103. It is to be understood that any conventional TTS system may be employed for implementing the present invention, although the preferred TTS system is International Business Machines' (IBM) trainable text-to-speech system disclosed in U.S. Patent Application Serial No. 09/084,679, entitled: "Methods For Generating Pitch And Duration Contours In A Text To Speech System," which is commonly assigned and incorporated herein by reference.

Briefly, with the IBM TTS system, the pronunciation of each word capable of being synthesized is characterized by its entry in a phonetic dictionary, with each entry comprising a string of phonemes which constitute the corresponding word. The TTS system concatenates segments of speech from phonemes in context to produce arbitrary sentences. A flat pitch equal to a training speaker's average pitch value is utilized to synthesize each segment. The duration of each segment is selected as the average duration of the segment in the training corpus plus a user-specified constant α times the standard deviation of the segment. The α term serves to control the rate of the synthesized speech and is fixed at a moderate value for all our experiments. The TTS system is built from data spoken by one male speaker who read 450 sentences of text. In operation, the IBM TTS system receives user-selected text sentence(s) and expands each word into a string of constituent phonemes by utilizing the synthesis dictionary.

Next, waveform segments for each phoneme are retrieved from storage and concatenated. The details of the procedure by which the waveform segments are chosen are described in the above-incorporated application. The pitch of the synthesis waveform is adjusted to flat using the pitch synchronous overlap and add (PSOLA) technique, which is also described in the above-incorporated application. The N synthetic waveforms are then saved to disk.

Each of the N synthetic waveforms 105 are input to the feature analysis module 101 and subjected to the same feature analysis as discussed above (for processing the original speech waveform) to generate N sets of feature vectors, with each set of feature vectors representing a corresponding one of the N synthetic waveforms 105. The N sets of feature vectors may be stored for subsequent processing. It is to be understood that for purposes of illustration and clarity, the system of Fig. 1 is shown as having two feature analysis modules 101, although the system is preferably implemented using one feature analysis module for processing both the original and synthetic waveforms.

A rescore module 106 compares the original waveform feature vectors with each of the N sets of synthetic waveform feature vectors and corresponding N-best text sequences to provide an N-best rescore output 110. In particular, this comparison processes begins in alignment module 107, whereby the original waveform feature vectors and each set of N synthetic waveform feature vectors are aligned to the text sequence of the corresponding N-best hypothesis. A distance computation module 108 calculates the distance between the original waveform and each of the N synthetic waveforms (using methods known to those skilled in the art). A comparator module 109 compares each of the calculated distances to rescore the N-best hypothesis based on the computed distances and determine the closest distance. The N-best text sequence corresponding to the closest synthetic waveform to the original speech is then output or otherwise saved as the final transcription of the utterance (i.e., the N-best rescore output 110).

Referring now to Figs. 2A and 2B, a flow diagram illustrates a preferred method for rescoring N-best hypotheses of an automatic speech recognition system in accordance with the present invention. Specifically, the flow diagram of Figs. 2A and 2B illustrates a detailed comparison process which is preferably employed in the rescore module 106 of Fig. 1. Initially, the rescore module 106 retrieves the original waveform feature vectors from memory (step 200). The comparison process is then initialized by setting a parameter N = 1 (where N represents the Nth-best hypothesis (text sequence) output from the speech recognition system 102) and setting a parameter "Best Distance" = infinity (where "Best Distance" is a threshold value that represents the smallest computed distance measure of previous iterations) (step 201).

Next, the Nth-best text sequence and the corresponding Nth synthetic waveform feature vectors are then retrieved from memory (step 202). The original waveform feature vectors and the Nth synthetic waveform feature vectors are then time-aligned to the Nth-best text sequence at the phoneme level (step 203). The alignment procedure preferably employs a Viterbi alignment process such as disclosed in "The Viterbi Algorithm," by G.D. Forney, Jr., Proc. IEEE, vol. 61, pp. 268-278, 1973. In particular, as is understood by those skilled in the art, the Viterbi alignment finds the most likely sequence of states given the acoustic observations, where each state is a sub-phonetic unit and the probability density function of the observations is modeled as a mixture of 60-dimensional Gaussians. It is to be appreciated that by time-aligning the original waveform and the Nth synthesized waveform to the Nth hypothesized text sequence at the phoneme level, each waveform may be segmented into contiguous time regions, with each region mapping to one phoneme in the phonetic expansion of the Nth text sequence (i.e., a segmentation of each waveform into phonemes).

After the alignment process, the mean of the feature vectors (frames) which align to each phoneme is computed for the original waveform and the Nth synthetic waveform (step 204). In this manner, the original waveform and the Nth synthetic waveform may be represented as a collection of mean feature vectors, with each mean feature vector representing the computed mean of all feature vectors aligning to a corresponding phoneme in the Nth text sequence. This process results in the generation of M mean feature vectors representing the original waveform and M mean feature vectors representing the Nth synthetic waveform (where M represents the number of phonemes in the expansion of the Nth text sequence into its constituent phonemes).

Next, a distance measure between each phoneme mean of the original waveform and the corresponding phoneme mean of the Nth synthetic waveform is computed (step 205). Although any suitable method may be employed for computing the distance measure, a Euclidean distance is preferably employed (by the distance computation module 108, Fig. 1). These individual distance measures (between each corresponding phoneme mean) are then summed to produce an overall distance measure (step 206) representing the "distance" between the original waveform and the Nth synthetic waveform corresponding to the Nth text sequence. Therefore, since the Nth synthetic waveform is derived from the Nth-best text sequence, it is to be appreciated that the overall distance measure indirectly represents the "distance" between the original waveform and the Nth-best text sequence.

A determination is then made as to whether the "distance" (which represents the overall distance between the original waveform and the Nth text sequence) is less than the current "Best Distance" value (step 207). If the "distance" is smaller than the "best distance" value (affirmative determination in step 207), a parameter "Best Text" is set so as to label the current Nth-best text sequence as the most accurate transcription encountered as compared to all previous iterations, and the parameter "best distance" is set equal to the current "distance" value (step 208).

A determination is then made as to whether there are any remaining N-best hypotheses for consideration (step 209). If there are additional N-best hypotheses (negative determination in step 209), the parameter N is incremented by one (step 210), and the next Nth-best text sequence and Nth synthetic waveform are retrieved from memory (return to step 202, Fig. 2A). This comparison process (steps 203-208) is repeated for N iterations (to rescore each N-best hypothesis). When it is determined that the final Nth-best hypothesis has been rescored (affirmative determination in step 209), the Nth-best text sequence having the minimum distance to the original waveform (as indicated by the "best text" and "best distance" parameters) is output (step 211). After the final output (step 211), the user may choose to rescore the N-best hypotheses of another original waveform (affirmative result in step 212) in which case the desired waveform will be retrieved from memory (return to step 200) and processed as described above. Alternatively, the user may terminate the rescore process and exit the program (step 213).

The above described preferred embodiment has been tested on speech degraded by the inclusion of additive noise in the form of background music. Test results have indicated an improvement of the word error rate from 27.8 percent to 27.3 percent using the two most-likely text hypotheses for each utterance. The improvement primarily results from a reduction in the number of erroneously inserted words.

It is to be appreciated by those skilled in the are that is some flexibility within the general framework of the present invention, thereby providing alternate embodiments of the above-described preferred embodiment. For instance, as noted above, different methods for measuring the distance between the original and synthetic waveforms may be substituted for the Euclidian distance measure described above.

In another embodiment of the present invention, in addition to re-ordering the N-best list based strictly on the distance of each synthesized hypothesis to the original waveform, the distance may be combined with other scores reflecting our confidence in the correctness of the N-th hypothesis, such as the likelihood of that hypothesis as assessed by the individual components comprising the automatic speech recognition system: the acoustic model and the language model. By combining the distance score with the scores from these sources, information provided by the decoder may be considered in conjunction with the new information provided by the distance score. For example, the scores may be combined by forming the following sum:

$$S_N = -D_N + (a \bullet A_N) + (b \bullet L_N)$$

where D_N is the distance of the N-th hypothesis from the original waveform (as described above); where A_N is the acoustic model score of the N-th hypothesis; where L_N is the language model score of the N-th hypothesis; and where a and b are constants. The text selected for output can then be the text associated with the N'-th hypothesis, where N' is the hypothesis whose score $S_{N'}$ is the maximum score among the N-best hypotheses considered.

In yet another embodiment, the original speech and/or synthetic speech may be further processed to compensate for speaker-dependent variations. For example, a vocal tract length normalization process (such as disclosed in "A Parametric Approach to Vocal-Tract-Length Normalization", by Eide et al., Proceedings of the Fifteenth Annual Speech Research Symposium, Johns Hopkins University, 1995; and "Speaker Normalization on Conversational Telephone Speech", by Wegmann et al., Vol. 1, Proc. ICASSP, pp. 339-341, 1996) may be performed on each test utterance to warp the frequency axis for each test speaker to match the vocal-tract characteristics of the speaker from whose data the TTS system was built. This would reduce the component in the distance between utterances due to differences between the speaker of the original test utterance and the speaker of the TTS system, which causes a relative increase of the contribution to the distance scores due to phonetic differences between the utterances.

Although illustrative embodiments have been described herein with reference to the accompanying drawings, it is to be understood that the present system and method is not limited to those precise embodiments, and that various other changes and modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention. All such changes and modifications are intended to be included within the scope of the invention as defined by the appended claims.

WHAT IS CLAIMED IS:

1. A program storage device readable by a machine, tangibly embodying a program of instructions executable by the machine to perform method steps for rescoring N-best hypotheses of a decoded original waveform output from an automatic speech recognition system, the N-best hypotheses comprising N text sequences, the method steps comprising:

generating a synthetic waveform for each of the N text sequences;

comparing each synthetic waveform with the original waveform to determine the synthetic waveform that is closest to the original waveform; and

selecting for output the text sequence corresponding to the synthetic waveform determined to be closest to the original waveform.

2. The program storage device of claim 1, wherein the instructions for performing the comparing step include instructions for performing the steps of:

aligning frames of the original waveform and frames of each synthetic waveform to a corresponding one of the N text sequences; and calculating the distance between the original waveform and each of the synthetic waveforms based on the corresponding alignments.

3. The program storage device of claim 2, wherein the instructions for performing the comparing step further include instructions for:

retrieving feature vectors corresponding to the original waveform; and

generating feature vectors for each synthetic waveform such that the feature vectors for the synthetic waveforms are similar in structure to the feature vectors of the original waveform;

wherein the alignment is performed by time-aligning the feature vectors of the original waveform and the feature vectors of each synthetic waveform with the corresponding one of the N text sequences.

4. The program storage device of claim 2, wherein the alignment is performed using Viterbi alignment process.

5. The program storage device of claim 2, wherein the alignment is performed on a phoneme level.

6. The program storage device of claim 2, wherein the instructions for calculating the distance include instructions for performing the steps of:

calculating an individual distance between each aligned frame of the original waveform and each of the N synthetic waveforms; and

summing the individual distances of the aligned frames of the original waveform and each synthetic waveform.

7. The program storage device of claim 1, wherein the instructions for performing the comparing step include instructions for performing the steps of:

(a) setting a parameter N=1;

(b) retrieving the Nth synthetic waveform and the corresponding Nth text sequence;

(c) time-aligning frames of the original waveform and frames of the Nth synthetic waveform to corresponding text of the Nth text sequence;

(d) computing an individual distance between each corresponding aligned frame of the original and Nth synthetic waveform;

(e) summing the individual distances to compute the distance between the original waveform and the Nth synthetic waveform;

(f) determining if the computed distance is less than a current best distance value;

(g) setting the current best distance value equal to the computed distance and saving the Nth text sequence for consideration as the final output, if the computed distance is determined to be less than the current best distance threshold;

(h) incrementing the parameter N by one; and

(i) repeating steps (b) through (h) until each of the N text sequences have been considered.

8. The program storage device of claim 7, wherein the instructions for performing the step of determining the individual distance (step d) include instructions for:

computing a mean feature vector of all feature vectors comprising each aligned frame for both the original and Nth synthetic waveform, wherein the individual distance for each aligned frame is calculated by determining a distance between each mean of the corresponding aligned frames.

9. A method for rescoring N-best hypotheses of a decoded original waveform output from an automatic speech

recognition system, the N-best hypotheses comprising N text sequences, the method comprising the steps of:

generating a synthetic waveform for each of the N text sequences;

comparing each synthetic waveform with the original waveform to determine the synthetic waveform that is closest to the original waveform; and

selecting for output the text sequence corresponding to the synthetic waveform determined to be closest to the original waveform.

10. The method of claim 9, wherein the comparing step includes the steps of:

aligning frames of the original waveform and frames of each synthetic waveform to a corresponding one of the N text sequences; and

calculating the distance between the original waveform and each of the synthetic waveforms based on the corresponding alignments.

11. The method of claim 10, wherein the comparing step further includes the steps of:

retrieving feature vectors corresponding to the original waveform; and generating feature vectors for each synthetic waveform such that the feature vectors for the synthetic waveforms are similar in structure to the feature vectors of the original waveform;

wherein the alignment is performed by time-aligning the feature vectors of the original waveform and the feature vectors of each synthetic waveform with the corresponding one of the N text sequences.

12. The method of claim 10, wherein the step of calculating the distance includes the steps of:

calculating an individual distance between each aligned frame of the original waveform and each of the N synthetic waveforms; and

summing the individual distances of the aligned frames of the original waveform and each synthetic waveform.

13. The method of claim 9, wherein the comparing step includes the steps of:

(a) setting a parameter N=1;

(b) retrieving the Nth synthetic waveform and the corresponding Nth text sequence;

(c) time-aligning frames of the original waveform and frames of the Nth synthetic waveform to corresponding text of the Nth text sequence;

(d) computing an individual distance between each corresponding aligned frame of the original and Nth synthetic waveform;

(e) summing the individual distances to compute the distance between the original waveform and the Nth synthetic waveform;

(f) determining if the computed distance is less than a current best distance value;

(g) setting the current best distance value equal to the computed distance and saving the Nth text sequence for consideration as the final output, if the computed distance is determined to be less than the current best distance threshold;

(h) incrementing the parameter N by one; and

(i) repeating steps (b) through (h) until each of the N text sequences have been considered.

14. The method of claim 13, wherein the step of determining the individual distance (step d) includes the steps of:

computing a mean feature vector of all feature vectors comprising each aligned frame for both the original and Nth synthetic waveform, wherein the individual distance for each aligned frame is calculated by determining a distance between each means of the corresponding aligned frames.

15. An automatic speech recognition system, comprising:

a decoder for decoding an original waveform of acoustic utterances to produce N text sequences, the N text sequences representing N-best hypotheses of the decoded original waveform;

a waveform generator for generating a synthetic waveform for each of the N text sequences; and

a comparator for comparing each synthetic waveform with the original waveform to rescore the N-best hypotheses.

16. The system of claim 15, further comprising a feature analysis processor adapted to generate a set of feature vectors for the original waveform and generate a set of feature vectors for each of the N synthetic waveforms using a similar feature analysis process.

17. The system of claim 15, further comprising a processor adapted to process one of the original waveform, the synthetic waveforms, and both, to compensate for speaker-dependent variations.

18. The system of claim 15, wherein the comparator comprises:

means for determining the synthetic waveform that is closest in distance to the original waveform; and

means for outputting the N text sequence corresponding to the synthetic waveform that is determined to be closest to the original waveform.

19. The system of claim 18, wherein the means for determining the closest synthetic waveform utilizes one of a distance score, a language model score, an acoustic model score, and a combination thereof, for determining the closest distance.

20. The system of claim 18, wherein the means for determining the closest synthetic waveform comprises:

means for aligning frames of the original waveform and frames of each synthetic waveform to a corresponding one of the N text sequences; and means for calculating the distance between the original waveform and each of the synthetic waveforms based on the corresponding alignments.

21. The system of claim 20, wherein the frames are aligned on a phoneme level.

22. The system of claim 20, wherein the means for calculating the distance comprises:

means for calculating an individual distance between each aligned frame of the original waveform and each of the N synthetic waveforms; and

means for summing the individual distances of the aligned frames of the original waveform and each synthetic waveform to compute the distance between the original waveform and each synthetic waveforms.

SYSTEM AND METHOD FOR RESCORING N-BEST HYPOTHESES OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM

ABSTRACT OF THE DISCLOSURE

A system and method for rescoring the N-best hypotheses from an automatic speech recognition system by comparing an original speech waveform to synthetic speech waveforms that are generated for each text sequence of the N-best hypotheses. A distance is calculated from the original speech waveform to each of the synthesized waveforms, and the text associated with the synthesized waveform that is determined to be closest to the original waveform is selected as the final hypothesis. The original waveform and each synthesized waveform are aligned to a corresponding text sequence on a phoneme level. The mean of the feature vectors which align to each phoneme is computed for the original waveform as well as for each of the synthesized hypotheses. The distance of a synthesized hypothesis to the original speech signal is then computed as the sum over all phonemes in the hypothesis of the Euclidean distance between the means of the feature vectors of the frames aligning to that phoneme for the original and the synthesized signals. The text of the hypothesis which is closest under the above metric to the original waveform is chosen as the final system output.

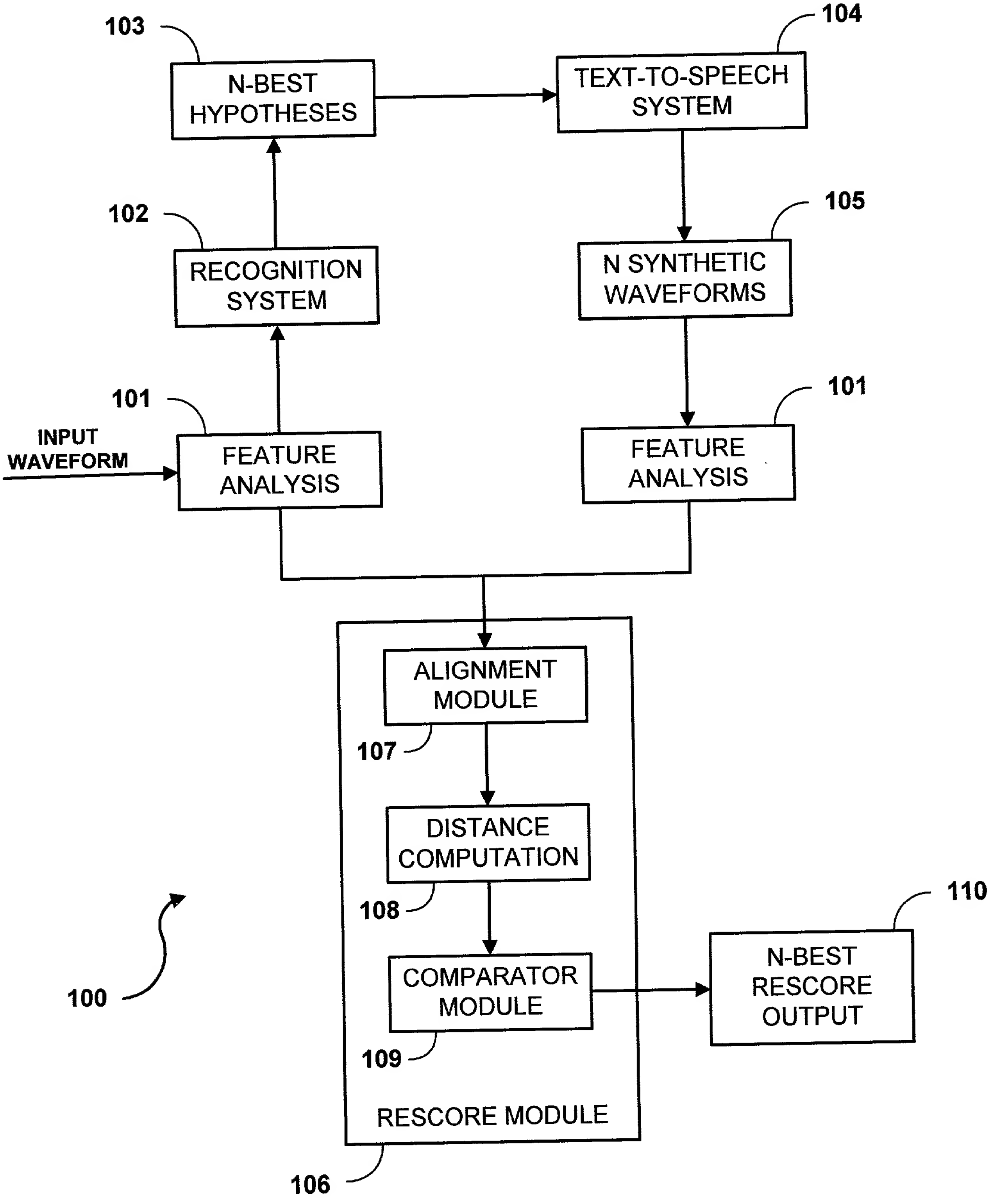
103
N-BEST HYPOTHESES
104
TEXT-TO-SPEECH SYSTEM
102
RECOGNITION SYSTEM
105
N SYNTHETIC WAVEFORMS
101
INPUT WAVEFORM
FEATURE ANALYSIS
101
FEATURE ANALYSIS
ALIGNMENT MODULE
107
DISTANCE COMPUTATION
108
110
COMPARATOR MODULE
N-BEST RESCORE OUTPUT
100
109
RESCORE MODULE
106

FIG. 1

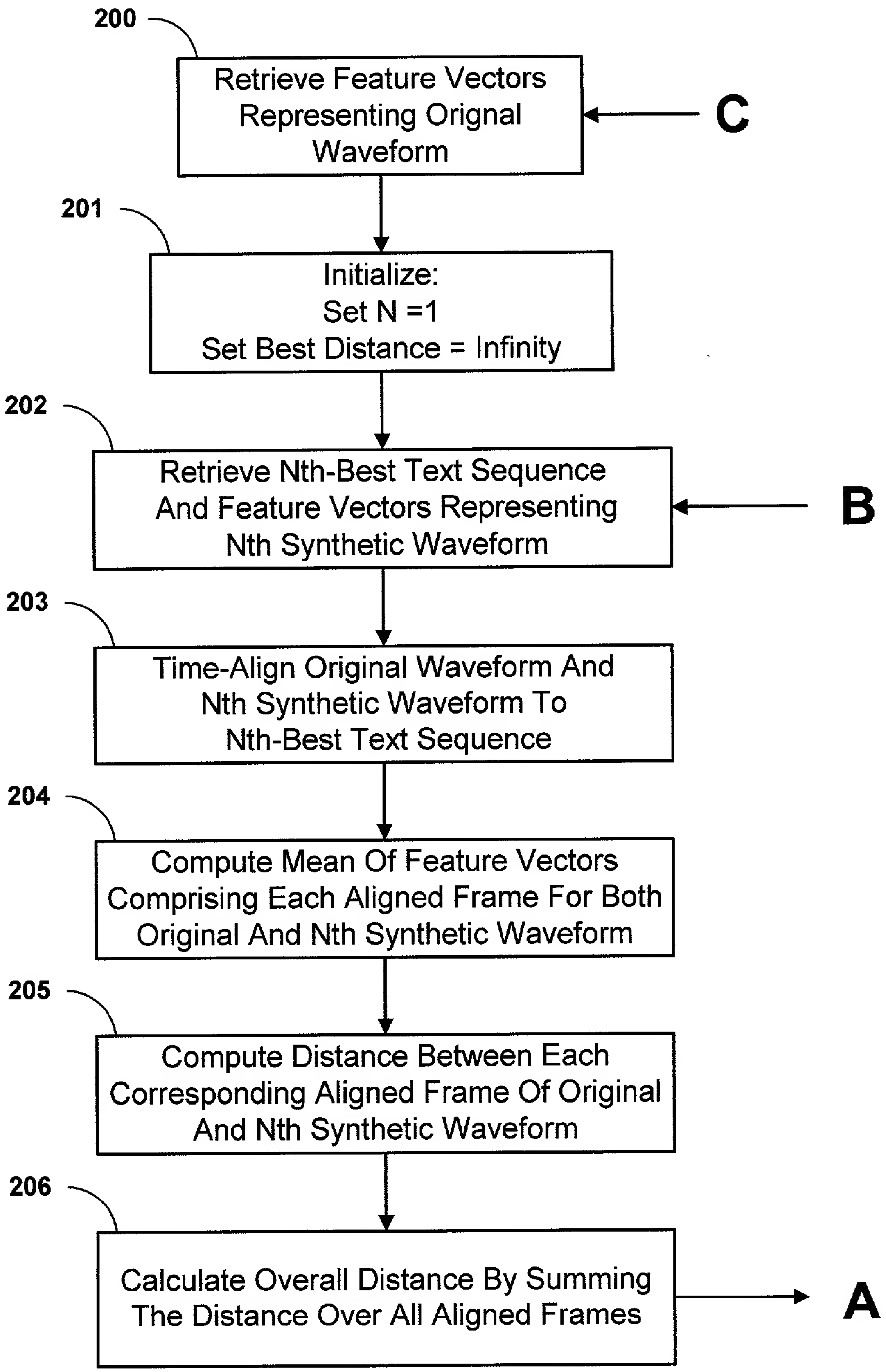
200
Retrieve Feature Vectors Representing Orignal Waveform
C
201
Initialize:
Set N =1
Set Best Distance = Infinity
202
Retrieve Nth-Best Text Sequence And Feature Vectors Representing Nth Synthetic Waveform
B
203
Time-Align Original Waveform And Nth Synthetic Waveform To Nth-Best Text Sequence
204
Compute Mean Of Feature Vectors Comprising Each Aligned Frame For Both Original And Nth Synthetic Waveform
205
Compute Distance Between Each Corresponding Aligned Frame Of Original And Nth Synthetic Waveform
206
Calculate Overall Distance By Summing The Distance Over All Aligned Frames
A

FIG. 2A

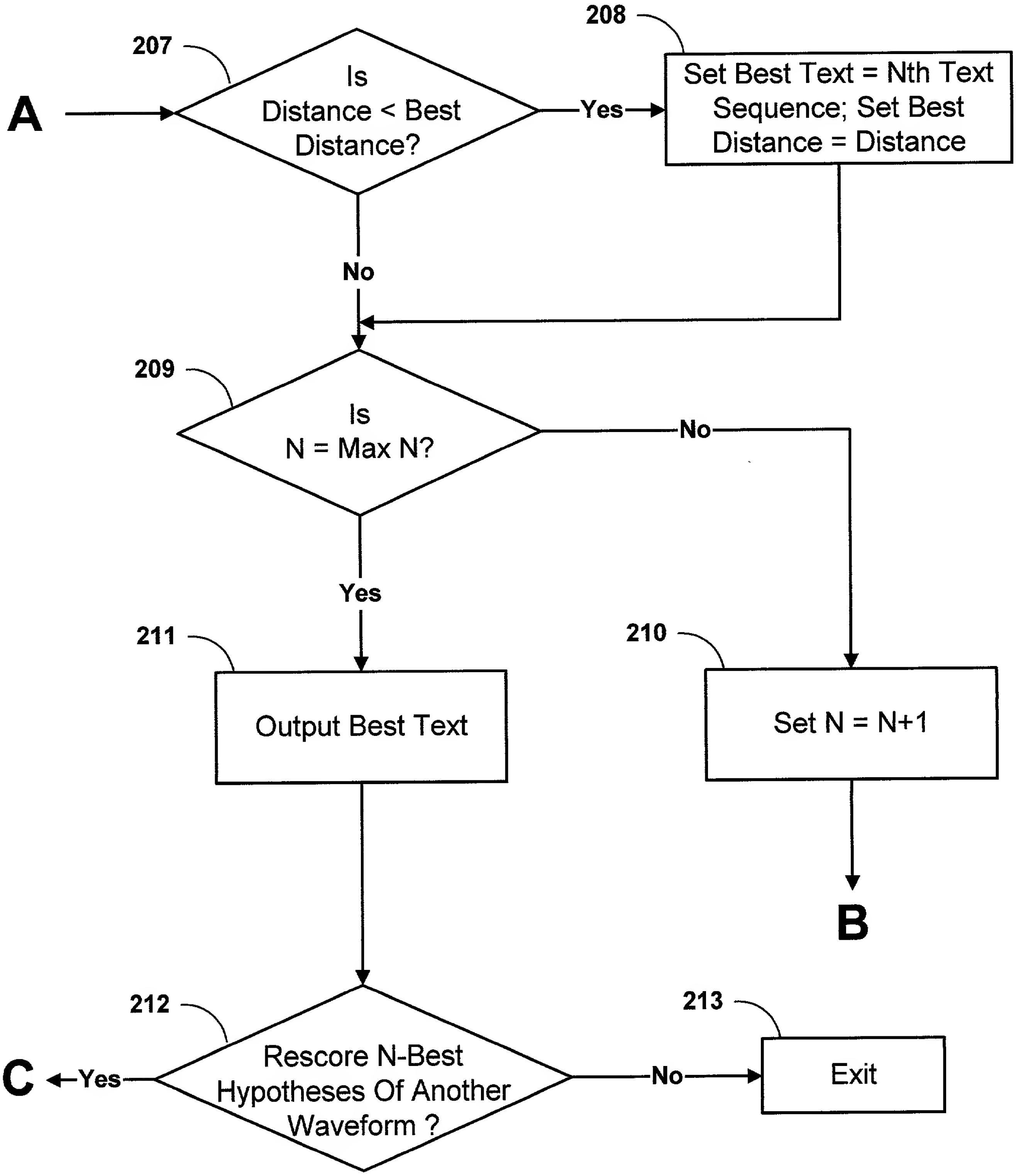
207
208
A
Is
Distance < Best
Distance?
Yes
Set Best Text = Nth Text
Sequence; Set Best
Distance = Distance
No
209
Is
N = Max N?
No
Yes
211
210
Output Best Text
Set N = N+1
B
212
213
C
Yes
Rescore N-Best
Hypotheses Of Another
Waveform ?
No
Exit

FIG. 2B

PTO/SB/01 (6/95) **DECLARATION** Attorney Docket No. Y0999-046 (8728-252)

AS A BELOW NAMED INVENTOR, I hereby declare that:

My residence, post office address and citizenship are as stated next to my name.

I believe that I am the original, first and sole *(if only one name is listed below)*, or an original, first and joint inventor *(if plural names are listed below)*, of the subject matter which is claimed and for which a patent is sought on the invention entitled:

TITLE: SYSTEM AND METHOD FOR RESCORING N-BEST HYPOTHESES OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM

the specification of which either is attached hereto or indicates an attorney docket no. ____________, or:

□ was filed in the U.S. Patent & Trademark Office on ________ and assigned Serial No. ______,

□ and *(if applicable)* was amended on ____,

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above. I acknowledge the duty to disclose information which is material to patentability and to the examination of this application in accordance with Title 37 of the Code of Federal Regulations §1.56. I hereby claim foreign priority benefits under Title 35, U.S. Code §119(a)-(d) or §365(b) of any foreign application(s) for patent or inventor's certificate, or §365(a) of any PCT international application which designated at least one country other than the United States, or §119(e) of any United States provisional application(s), listed below and have also identified below any foreign applications for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

Priority Claimed:

____ Yes [ ] No [ ]

(Application Number) *(Country)* *(Day/Month/Year filed)*

____ Yes [ ] No [ ]

(Application Number) *(Country)* *(Day/Month/Year filed)*

I hereby claim the benefit under Title 35, U.S. Code, §120, of any United States application(s), or §365(c) of any PCT International application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT International application(s) in the manner provided by the first paragraph of Title 35, U.S. Code, §112, I acknowledge the duty to disclose information material to patentability as defined in Title 37, The Code of Federal Regulations, §1.56(a) which became available between the filing date of the prior application and the national or PCT international filing date of this application:

____

(Application Serial Number) *(Filing Date)* *(STATUS: patented, pending, abandoned)*

____

(Application Serial Number) *(Filing Date)* *(STATUS: patented, pending, abandoned)*

I hereby appoint the following attorneys: **MANNY W. SCHECTER**, Reg. No. 31,722; **TERRY J. ILARDI**, Reg. 29,936; **CHRISTOPHER A. HUGHES**, Reg. No. 26,914; **EDWARD A. PENNINGTON**, Reg. No. 32,588; **JOHN E. HOEL**, Reg. No. 26,279; **JOSEPH C. REDMOND, Jr.**, Reg. No. 18,753; **KEVIN M. JORDAN**, Reg. No. 40,277; **STEPHEN C. KAUFMAN**, Reg. No. 29,551; **JAY P. SBROLLINI**, Reg. No. 36,266; **DAVID M. SHOFI**, Reg. No. 39,835; **ROBERT M. TREPP**, Reg. No. 25,933; **LOUIS P. HERZBERG**, Reg. No. 41,500; **DANIEL P. MORRIS**, Reg. No. 32,053; **PAUL J. OTTERSTEDT**, Reg. No. 37,411; and **DOUGLAS W. CAMERON**, Reg. No. 31,596; each of them of **INTERNATIONAL BUSINESS MACHINES CORPORATION**, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598; to prosecute this application and to transact all business in the U.S. Patent and Trademark Office connected therewith and with any divisional, continuation, continuation-in-part, reissue or re-examination application, with full power of appointment and with full power to substitute an associate attorney or agent, and to receive all patents which may issue thereon, and request that all correspondence be addressed to:

Frank Chau, Esq.
F. CHAU & ASSOCIATES, LLP
1900 Hempstead Turnpike, Suite 501
East Meadow, New York 11554
Tel.: 516-357-0091

I HEREBY DECLARE that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under §1001 of Title 18 U.S. Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

FULL NAME OF FIRST OR SOLE INVENTOR: Raimo Bakis Citizenship U.S.A.

Inventor's signature: ______ Date: 4/1/99
Residence & Post Office Address: 28 Winterberry Lane
Briarcliff Manor, NY 10510

FULL NAME OF SECOND JOINT INVENTOR: Ellen M. Eide Citizenship U.S.A.

Inventor's signature: ______ Date: April 1, 1999
Residence & Post Office Address: 603 Kensington Way
Mount Kisco, NY 10549

FULL NAME OF THIRD JOINT INVENTOR: ______ Citizenship ______

Inventor's signature: ______ Date: ______
Residence & Post Office Address:

FULL NAME OF FOURTH JOINT INVENTOR: ______ Citizenship ______

Inventor's signature: ______ Date: ______
Residence & Post Office Address:

FULL NAME OF FIFTH JOINT INVENTOR: ______ Citizenship ______

Inventor's signature: ______ Date: ______
Residence & Post Office Address:

FULL NAME OF SIXTH JOINT INVENTOR: ______ Citizenship ______

Inventor's signature: ______ Date: ______
Residence & Post Office Address:

Attorney Docket No. YO999-046 (8728-252)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANT(S): Raimo Bakis, Ellen M. Eide

SERIAL NO.: Unassigned

FILED: Concurrently herewith

FOR: SYSTEM AND METHOD FOR RESCORING N-BEST HYPOTHESES OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM

ASSOCIATE POWER OF ATTORNEY

Please recognize **FRANK CHAU**, Reg. No. 34,136; **JAMES J. BITETTO**, Reg. No. 40,513; **FRANK V. DeROSA**, Reg. No. 43,584; and **GASPARE J. RANDAZZO**, Reg. No. 41,528; each of them of **F. CHAU & ASSOCIATES, LLP**, 1900 Hempstead Turnpike, Suite 501, East Meadow, New York 11554 as associate attorneys in the above-mentioned application, with full power to prosecute said application, to make alterations and amendments therein, and to transact all business in the Patent and Trademark Office connected therewith.

Telephone calls should be made to Frank Chau by dialing (516) 357-0091.

All written communications are to be sent to Frank Chau, Esq., F. Chau & Associates, LLP, 1900 Hempstead Turnpike, Suite 501, East Meadow, New York 11554.

Paul J. Otterstedt

Manny W. Schecter
Registration No. 31,722
Paul J. Otterstedt
Registration No. 37,411
Attorney for Applicant(s)

International Business Machines
Corporation
T.J. Watson Research Center
Route 134 and Kitchawan Road
Yorktown Heights, New York 10598